PRIVACY IN AI/ML

Data → $\boxed{\text{ML/AI Alg}}$ → Model → WORLD

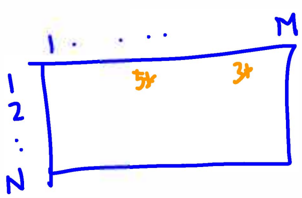
Health care records

want to be private.

Predicts disease = Y/N

Anonymisation

Netflix



- CRYPTOGRAPHIC SOLUTION ✗
- $\boxed{\text{STATISTICAL}}$ SOLUTION TO PRIVACY ✓

Note Title

Data: $\{\underbrace{x_1, \ldots, x_n}_{\text{Truth}}\}$ $x_i \in \{0, 1\}$

1 → voted.

0 → did not vote

Randomly toss coin	Reveal truth
PRIVACY ✓	PRIVACY ✗
UTILITY ✗	UTILITY ✓

RANDOMIZED RESPONSE

Note Title

Truth $\{x_1, \dots, x_n\}$ $x_i \in \{0,1\}$

Revealed $Y_i = \begin{cases} x_i & \text{with probability } \dfrac{e^{\epsilon}}{1+e^{\epsilon}} \quad \text{TRUTH} \\ 1 - x_i & \text{with probability } \dfrac{1}{1+e^{\epsilon}} \quad \text{FALSEHOOD} \end{cases}$

Consider two datasets

$X = \{x_1, \dots, x_n\}$ $X' = \{x_1', \dots, x_n'\}$

Assume $x_i = x_i'$ $\forall i \in \{1, \dots, n-1\}$

Note Title

Revealed

$RR(x) = y = [y_1, \ldots, y_n]$ $\qquad y_i \in \{0,1\}$

$RR(x') = y' = [y_1', \ldots, y_n']$ $\qquad y_i' \in \{0,1\}$

$$\text{Prob}\left(RR(x) = b\right) = \text{Prob}\left([y_1, y_2 \cdots y_n] = [b_1, b_2 \cdots, b_n]\right)$$

b: $[b_1\ b_2 \cdots b_n]$

eg: $[0\ 1\ 0\ 1\ 1]$

$$= \prod_{i=1}^{n} P(y_i = b_i)$$ [Independence of $y_1, \cdots, y_n$]

$$= \left[\prod_{i=1}^{n-1} P(y_i = b_i)\right] P(y_n = b_n) \quad \text{— (1)}$$

Note Title

$$\text{Prob}\left(RR(x') = b\right) = \prod_{i=1}^{n} P\left(y_i' = b_i\right)$$

$$= \underbrace{\left(\prod_{i=1}^{n-1} P\left(y_i' = b_i\right)\right)} P\left(y_n' = b_n\right)$$

$$= \left(\prod_{i=1}^{n} P\left(y_i = b_i\right)\right) \underline{P\left(y_n' = b_n\right)} \quad \text{— } (2)$$

$$\checkmark \; \frac{P\left(y_n = b_n\right)}{\checkmark \; P\left(y_n' = b_n\right)} = \begin{cases} \dfrac{e^{\epsilon}/(1+e^{\epsilon})}{1/(1+e^{\epsilon})} = e^{\epsilon} & \underline{b_n = x_n} \; \checkmark \\[2ex] \dfrac{1/(1+e^{\epsilon})}{e^{\epsilon}/(1+e^{\epsilon})} = \underline{e^{-\epsilon}} & b_n \neq x_n \; \left(\text{i.e., } b_n = 1 - x_n\right) \; \checkmark \end{cases}$$

Note Title

In both cases,

$$\frac{P(y_n = b_n)}{P(y_n' = b_n)} \le e^{\epsilon}$$

$$\Rightarrow \boxed{P(y_n = b_n) \le e^{\epsilon} \cdot P(y_n' = b_n)} \quad \text{— (3)}$$

$$\frac{Pr(RR(x) = b)}{Pr(RR(x') = b)} \underset{(1) \,\&\, (2)}{=} \frac{P(y_n = b_n)}{P(y_n' = b_n)} \underset{(3)}{\le} e^{\epsilon}$$

Note Title

$$\Pr(RR(x) = b) \leq e^{\epsilon} \Pr(RR(x') = b)$$

observe

$$\Rightarrow \boxed{\Pr(RR(x') = b) \cdot e^{-\epsilon} \leq \Pr(RR(x) = b) \leq e^{\epsilon} \Pr(RR(x') = b)}$$

UTILITY of the mechanism

Note Title

want $\frac{1}{n}\sum_{i=1}^{n} x_i$

$x_i \xrightarrow{\text{Randomization}} y_i$

$\{0,1\}$

$$\underline{\mathbb{E}[y_i]} = \left(\frac{e^{\epsilon}}{1+e^{\epsilon}}\right) x_i + \left(\frac{1}{1+e^{\epsilon}}\right)(1-x_i)$$

$$\frac{e^{\epsilon} x_i + 1 - x_i}{1+e^{\epsilon}} = \boxed{x_i\left(\frac{e^{\epsilon}-1}{e^{\epsilon}+1}\right) + \frac{1}{1+e^{\epsilon}}}$$

Note Title

unbiasing.

$$X_i \rightarrow Y_i \xrightarrow{\text{unbiasing}} Z_i$$

$$Z_i = \left(Y_i - \frac{1}{1+e^{\epsilon}}\right) \cdot \left(\frac{e^{\epsilon}+1}{e^{\epsilon}-1}\right)$$

$$\mathbb{E}[Z_i] = X_i \quad [\text{Exercise}]$$

$$\begin{array}{ccc} X_1, & \cdots & X_n \\ \downarrow & & \downarrow \\ Y_1 & \cdots & Y_n \\ \downarrow & & \downarrow \\ Z_1 & \cdots & Z_n \end{array}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Guess for $\bar{X} = \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$

$$\boxed{\mathbb{E}[\bar{Z}] = ? \quad = \bar{X}}$$

Note Title

Utility : $\left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} z_i} \right|$

with high probability

$\leq O\left(\frac{1}{\epsilon\sqrt{n}}\right)$ ← UTILITY

ϵ Controls Privacy - utility tradeoff

TRUSTED CURATOR

Data $\rightarrow$ Curator $\rightarrow$ world

Data $\rightarrow$ Curator $\rightarrow$ world.

DIFFERENTIAL PRIVACY $\rightarrow$ Cynthia Dwork

Note Title

$\{x_1, \ldots, x_n\}$ $\quad x_i \in \{0,1\}$

$RR(x) = \{y_1, \ldots, y_n\}$

$y_i = x_i$ w.p $\frac{e^{\epsilon}}{1+e^{\epsilon}}$

$= 1-x_i$ $\quad 1/_{1+e^{\epsilon}}$

x, x'

Fix any $b \in \{0,1\}^n$

$$\forall b \in \{0,1\}^n \quad \frac{\Pr(RR(x) = b)}{\Pr(RR(x') = b)} \leq e^{\epsilon}$$

Privacy

Utility

$x_1, \ldots, x_n$

$\downarrow \qquad \downarrow$

$y_1, \ldots, y_n$

$\downarrow \qquad \downarrow$

$z_1, \ldots, z_n$

$\hookrightarrow \frac{1}{n}\sum_{i=1}^{n} z_i$

w.h.p

$\left|\frac{1}{n}\sum x_i - \frac{1}{n}\sum z_i\right|$

$\leq O\left(\frac{1}{\epsilon\sqrt{n}}\right)$

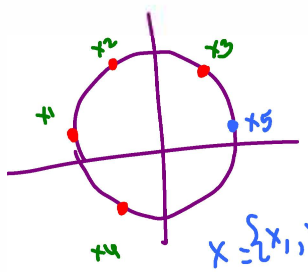
Note Title

Trusted curator Model

Differential Privacy [Cynthia Dwork]

Let $M: \mathcal{X}^n \to \mathcal{Y}$. Consider two "neighboring" datasets x and $x' \in \mathcal{X}^n$

$\{0,1\}^n \to \{0,1\}^n$



$X = \{x_1, x_2, x_3, x_4\}$

$X' = \{x_1, x_2, x_3, x_5\}$

M is ϵ-differentially private if for all x, x' neighbouring and all $S \subseteq \mathcal{Y}$

Note Title

$$\boxed{\frac{\Pr(M(x) \in S)}{\Pr(M(x') \in S)} \le e^{\epsilon}}$$

In the previous example

$\mathcal{Y} = \{0,1\}^n$ $\qquad S \subseteq \{0,1\}^n$ $\qquad S = \{b_1, b_{25}, b_{108}\}$

$$\frac{\Pr(M(x) \subseteq S)}{\Pr(M(x') \subseteq S)} = \frac{\Pr(M(x) = b_1) + \Pr(M(x) = b_{25}) + \Pr(M(x) = b_{108})}{\Pr(M(x') = b_1) + \cdots}$$

Note Title

LAPLACE MECHANISM

Sensitivity

$f : \mathcal{X}^n \to \mathbb{R}$ (average)

$\Delta = \max_{\substack{x, x' \\ \text{neighbouring}}} \left| f(x) - f(x') \right|$

x, x'

$\left| \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} x_i' \right|$

$\frac{1}{n} \left| \cancel{\sum_{i=1}^{n-1} x_i} + x_n - \cancel{\sum_{i=1}^{n} x_i'} - x_n' \right|$

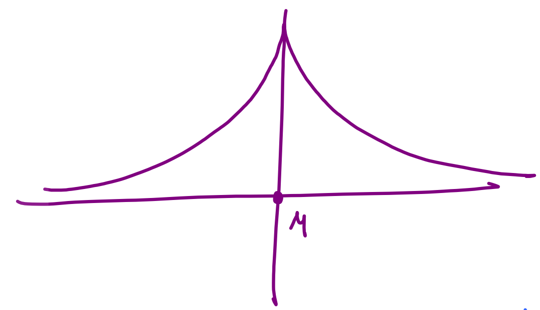
$= \frac{1}{n} \left| x_n - x_n' \right|$

$\Delta = \frac{1}{n}$

Note Title

LAPLACE DISTRIBUTION

$$f_{Lap}(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$



Laplace mechanism : $f(x) + \eta$

noise

Laplace $\left(0, \frac{\Delta}{\epsilon}\right)$

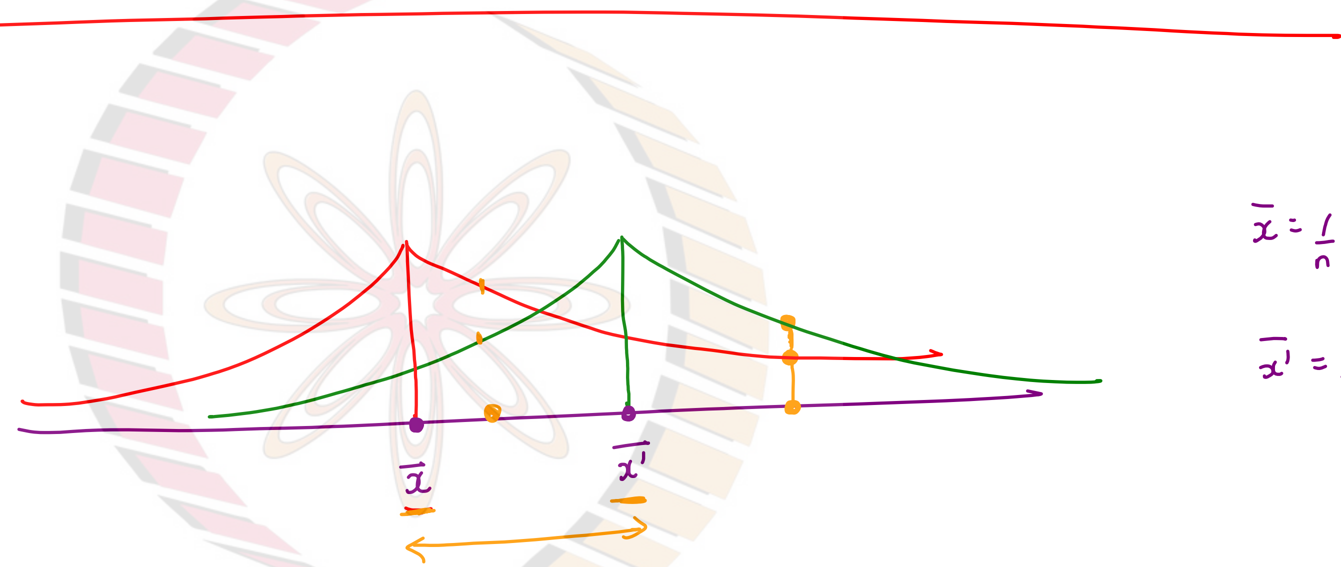
Sensitivity of f

Privacy parameter

average $+ \eta$ $\quad \eta \sim Lap\left(0, \frac{1}{n\epsilon}\right)$

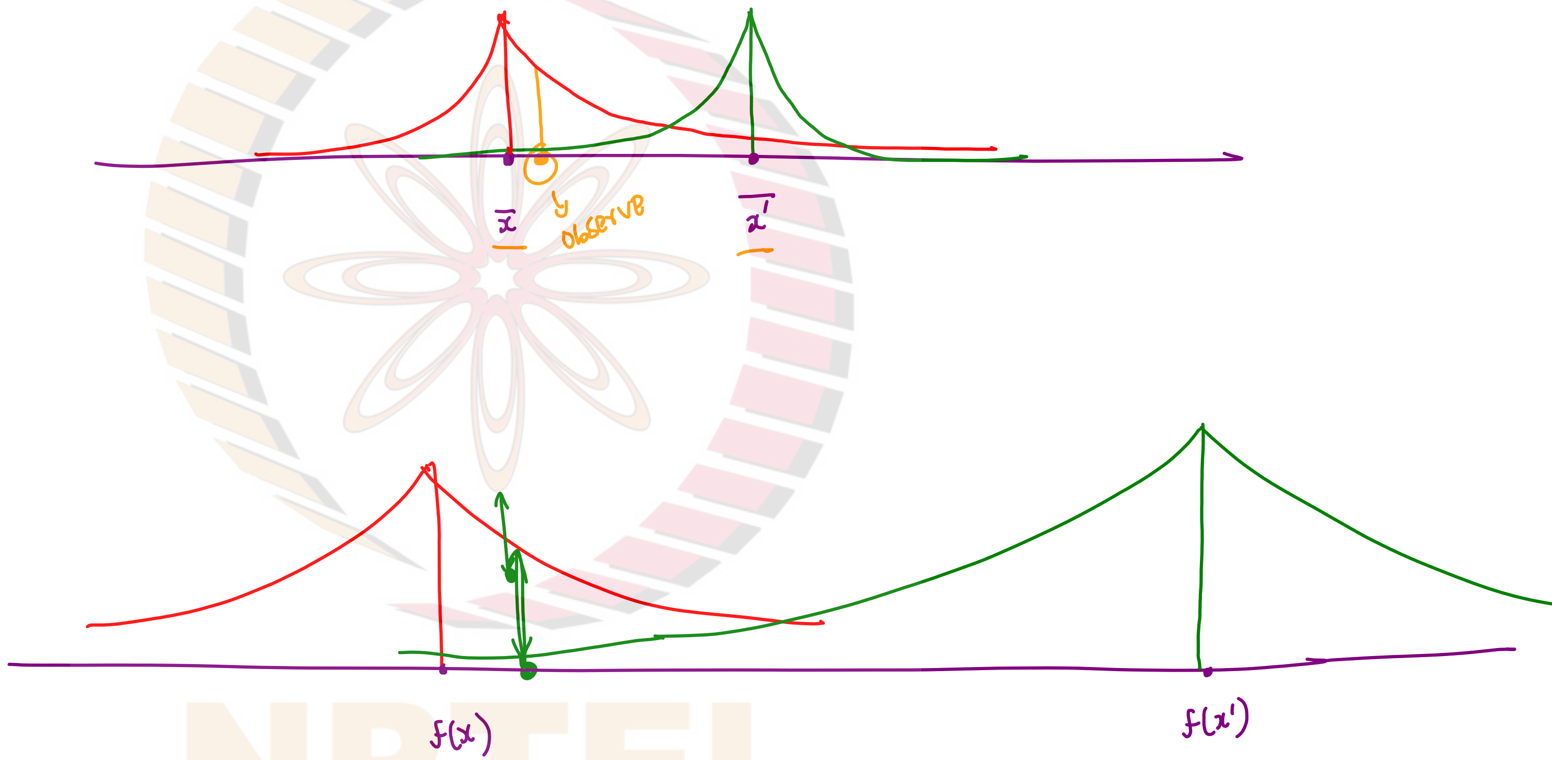
Note Title

- Can argue that the Laplace mechanism is $\underline{\epsilon\text{-DP}}$



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i'$$


$\bar{x}$
observe
$\bar{x}'$
$f(x)$
$f(x')$

Utility

w.h.p $\left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{Truth}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i + \eta \right)}_{\text{actual released value}} \right|$

deviation from truth.

actual released value $f(x) + \eta$

LAPLACE	R.R
$O\left(\frac{1}{\epsilon n}\right)$	$O\left(\frac{1}{\epsilon \sqrt{n}}\right)$